LETTER Special Section on Robust Speech Processing in Realistic Environments

Local Peak Enhancement for In-Car Speech Recognition in Noisy Environment

Osamu ICHIKAWA[†a)], Takashi FUKUDA[†], *and* Masafumi NISHIMURA[†], *Members*

SUMMARY The accuracy of automatic speech recognition in a car is significantly degraded in a very low SNR (Signal to Noise Ratio) situation such as "Fan high" or "Window open". In such cases, speech signals are often buried in broadband noise. Although several existing noise reduction algorithms are known to improve the accuracy, other approaches that can work with them are still required for further improvement. One of the candidates is enhancement of the harmonic structures in human voices. However, most conventional approaches are based on comb filtering, and it is difficult to use them in practical situations, because their assumptions for F0 detection and for voiced/unvoiced detection are not accurate enough in realistic noisy environments. In this paper, we propose a new approach that does not rely on such detection. An observed power spectrum is directly converted into a filter for speech enhancement, by retaining only the local peaks considered to be harmonic structures in the human voice. In our experiments, this approach reduced the word error rate by 17% in realistic automobile environments. Also, it showed further improvement when used with existing noise reduction methods.

key words: *harmonics, formant, speech enhancement, noise reduction, speech recognition*

1. Introduction

Automatic speech recognition in a car shows significant degradation in the following cases:

1. Simultaneous voices from passengers
2. Sounds coming from a car radio, TV, or CD player
3. Very low SNR situations such as "Fan high" or "Window open"

For Cases 1 and 2, beam former [1] and echo canceller [2] technologies are expected to solve the problems. However in Case 3, the speech signals are often buried in noise, and it is difficult to obtain sufficient recovery only with existing noise reduction algorithms such as a Wiener Filter [3] or Spectral Subtraction (SS) [4]. Therefore, for further improvement, different approaches beyond reducing noise should be combined with existing noise reduction algorithms.

One of the candidates is enhancements of the harmonic structures in human voices. Comb filtering [5] and its variants [6] were proposed and showed good performance, especially in mixed speech cases. However, it is not commonly integrated into commercial ASR products, and especially not for automobiles. This is because designing the comb filter relies on the accurate estimation of F0 (the fundamental frequency) and the accurate discrimination between voiced and unvoiced speech. It was reported that errors at this stage have detrimental effects on the performance [7]. Szymanski et al. proposed Comb Filter Decomposition [8] that does not require F0 estimation, but their experiment was limited to white Gaussian noise. Also, they used comb filtering in the time domain, which does not allow existing noise reduction algorithms to preprocess the input for the comb filter in the spectral domain.

Another candidate would be a matching algorithm to put larger weights on frequencies having larger spectral powers as the decoder calculates likelihoods [9], [10]. This is based on the assumption that frequencies having more spectral power are noise robust and most likely to be the formant frequencies in voiced speech frames. Huang et al. enhanced the logic for the MFCC domain [11], but this involved adding autocorrelation into their decoding process.

In this paper, we propose a new approach for the speech enhancement. It uses a filter designed to enhance the harmonic structure which is observed as local peaks at regular distances in the spectrum domain. It does not depend on F0 or voiced/unvoiced detection. Since it works as a front-end for both training and decoding, it does not require any changes in existing decoders. This new method will be referred to as LPE (Local Peak Enhancement) in the following sections.

2. Proposed Method

2.1 LPE

Figure 1 shows the whole process of LPE and sample outputs at each step for both a voiced frame and a noise frame. The process is the same for entire frames, but the generated filter looks very different depending on whether or not the frame is voiced speech, as shown in the figure.

In the first step, an observed spectrum $y_T(j)$ is converted to a log power spectrum $Y_T(j)$.

$$Y_T(j) = \log(y_T(j)) \tag{1}$$

Here, the index T is a frame number and j is the bin number of the DFT corresponding to the subband frequency. The process described in this section should be performed for each T.

Then the log power spectrum is converted to a cepstrum $C_T(i)$ by using $D(i, j)$, a DCT (Discrete Cosine Transformation) matrix.

Manuscript received July 4, 2007.
Manuscript revised September 3, 2007.

[†]The authors are with Tokyo Research Laboratory, IBM Japan Ltd., Yamato-shi, 242–8502 Japan.
a) E-mail: ICHIKAW@jp.ibm.com

DOI: 10.1093/ietisy/e91–d.3.635

$$C_T(i) = \sum_j D(i,j) \cdot Y_T(j) \tag{2}$$

The cepstra represent the curvatures of the log power spectra. The lower cepstra correspond to long oscillations, and the upper cepstra correspond to short oscillations. We need only the medium oscillations. The range of the cepstra is chosen to cover possible harmonic structures in the human voice. Therefore the lower and the upper cepstra should be filtered out.

$$\begin{aligned} \hat{C}_T(i) &= epsilon \cdot C_T(i) && \text{if } i < lower\_cep \\ & && \text{or } i > upper\_cep \\ \hat{C}_T(i) &= C_T(i) && \text{else} \end{aligned} \tag{3}$$

In our experiments, $lower\_cep = 40$ and $upper\_cep = 160$ for a 16 KHz sampling frequency with an FFT lcngth of 512 samples. This corresponds to an F0 range from 100 Hz to 400 Hz for the human voice, with *epsilon* being close to zero. We set it to 10^{-3}.

The filtered cepstrum $\hat{C}_T(i)$ is converted back to a log power spectrum by using an I-DCT.

$$W_T(j) = \sum_i D^{-1}(j,i) \cdot \hat{C}_T(i) \tag{4}$$

Then it is converted back to a linear power spectrum, and it is normalized so that the average is 1.0.

$$w_T(j) = \exp(W_T(j)) \tag{5}$$

$$\bar{w}_T(j) = w_T(j) \cdot \frac{Num\_bin}{\sum_k^{Num\_bin} w_T(k)} \tag{6}$$

Here, *Num_bin* is the number of bins used in the FFT. The filter is obtained as $\bar{w}_T(j)$. Finally, the enhanced output $z_T(j)$ is obtained as

$$z_T(j) = y_T(j) \cdot \bar{w}_T(j) \tag{7}$$

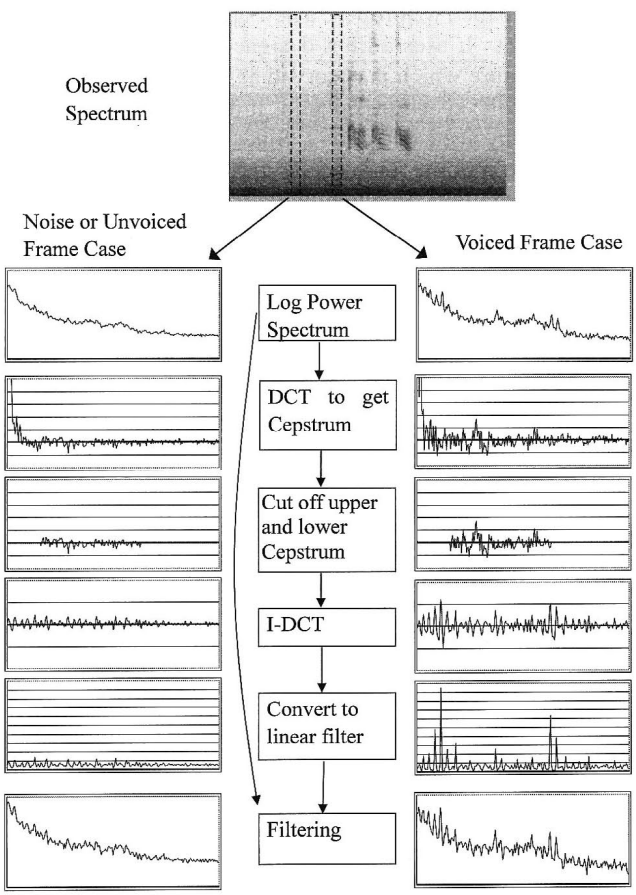
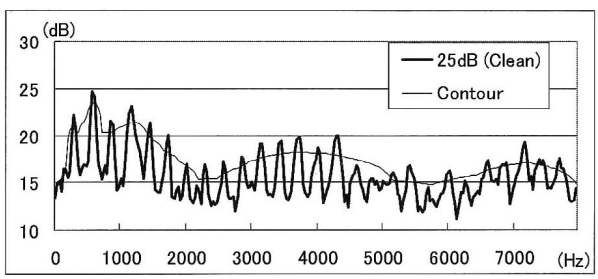


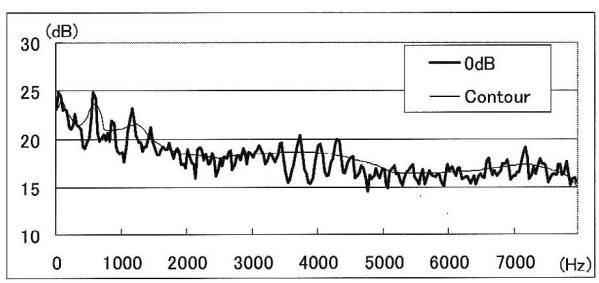
Fig. 1 Process of LPE.

2.2 Characteristics of an LPE Filter

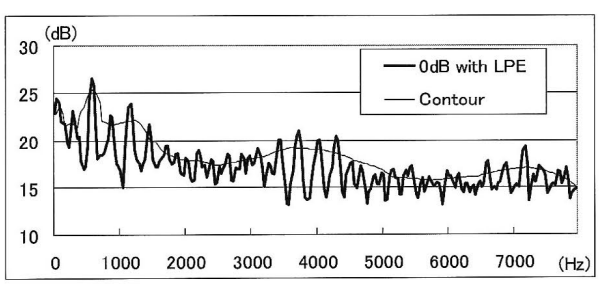
As shown in Fig. 1, the filter of LPE is made directly from the observed spectrum. Therefore, F0 estimation is not required. For a noise frame or an unvoiced speech frame, it will be designed to be almost flat. This means LPE does almost nothing to such frames, and therefore, LPE does not require voiced/unvoiced detection.



(a) Original sound



(b) Fan noise overlapped at SNR 0 dB



(c) Fan noise overlapped at SNR 0 dB and processed by LPE

Fig. 2 Spectrums of vowel /u/ recorded in a stationary car with and without fan noise overlapped at the specified SNR. The spectrum contour is plotted with Mel-Filtering.

For voiced speech frames, the LPE filter is designed to enhance the harmonic structure in the observed spectrum. Unlike a comb filter, the LPE filter is not uniform over all frequencies. It is more focused on the frequencies where harmonic structures are observed in the input spectrum. Therefore the acoustic model should be retrained with LPE for automatic speech recognition.

Figure 2 shows how a spectrum is degraded by a noise. In Fig. 2 (a), the original clean spectrum shows three formants around 600 Hz, 1200 Hz, and 3500 Hz. However, in Fig. 2 (b), they are less conspicuous, and the spectrum contour is flat. In contrast, LPE retains more of the characteristics of the formants, as shown in Fig. 2 (c).

Harmonic structures are conspicuous around frequencies having larger spectral powers in the voiced speech frames, and they are most likely to be formant frequencies. Therefore, this approach inherently involves formant enhancement as well as harmonic enhancement, under the assumption that the noise has a broad spectrum and the harmonic structure is not locally destroyed by the noise.

3. Experiments

3.1 Testing Data

We used CENSREC-3, an evaluation framework for isolated Japanese word recognition in actual moving-automobile environments. This data was collected by IPSJ, and is widely used to evaluate noise reduction algorithms [12]. It has speech data both for training and testing for automatic speech recognition using matched acoustic models.

The test data in the database was recorded under 16 environmental conditions using combinations of three vehicle speeds and six kinds of in-car environments as shown in Table 1. A total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) were recorded at a 16 KHz sampling frequency.

For training, each driver's speech saying phonetically balanced sentences was recorded under two conditions: while idling and while driving on a city street in a normal in-car environment. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded with a close-talking microphone and a hands-free microphone.

In this experiment, we used only hands-free microphone data for both training and testing. The acoustic models were trained with both idling data and driving data for the front-end processing being tested. This corresponds to Condition 3 as defined in CENSREC-3. The evaluation category is zero, which means no changes at the backend.

3.2 Conventional Methods

Comb-filtering needs F0 estimation and voiced/unvoiced detection. We used the "Pitch command" in SPTK-3.0 [13] to obtain this information. We used a low-end frequency of 100 Hz and an upper frequency limit of 400 Hz, so to be compatible with LPE experiment. The voiced/unvoiced threshold was empirically set to 7.0, because it gave us a better result than the SPTK default value. Figure 3 shows an example of F0 information by SPTK. We see many outliers in the low SNR conditions. Also, the vowels in the last part of the sentence were not recognized as voiced sounds. Based on the F0 and voiced/unvoiced information, the comb

Table 1 Word accuracy and estimated SNRs according to the environmental conditions. SNR was calculated for the baseline data after a 250 Hz high-pass filtering.

CENSREC-3 (Condition 3)			SNR (dB)	Word Accuracy (%) Base Line	Comb Filter	LPE
Idling	Audio off	Normal	16.2	99.7	98.8	99.7
		Hazard on	15.3	98.7	95.3	96.8
		Fan low	11.3	94.6	87.7	94.8
		Fan high	6.2	53.4	55.0	60.3
		Window open	10.5	90.0	85.4	92.7
	Audio on		9.9	81.4	73.2	56.4
Low speed	Audio off	Normal	10.9	99.3	96.6	98.7
		Fan low	9.7	95.1	91.8	94.7
		Fan high	6.7	62.7	66.2	69.1
		Window open	9.3	66.2	70.6	74.3
	Audio on		6.7	79.0	74.7	61.6
High speed	Audio off	Normal	7.5	95.0	94.3	96.2
		Fan low	7.1	89.0	86.7	89.7
		Fan high	6.1	58.2	62.1	63.6
		Window open	7.2	22.2	35.8	40.4
	Audio on		3.9	79.3	69.0	66.6
Average (ALL)				78.9	77.6	78.4
Average (Audio off)				**78.8**	**78.9**	**82.4**
Average (Audio on)				79.9	72.3	61.5
Average (Fan high)				58.1	61.1	64.3
Average (Window open)				59.5	63.9	69.1

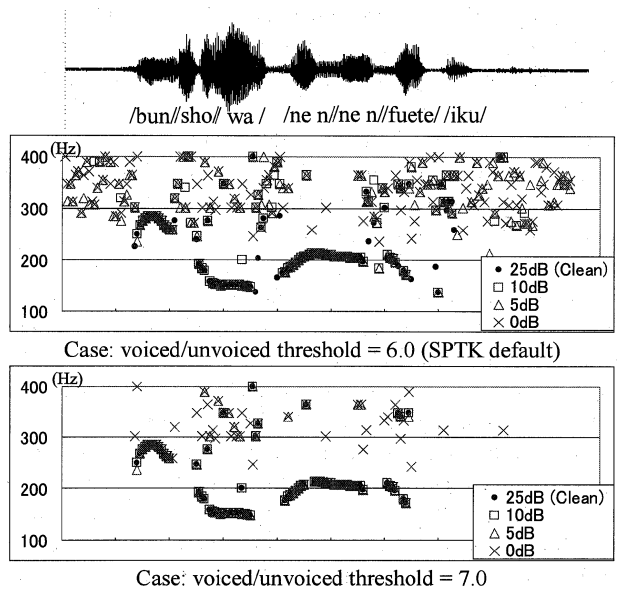


Fig. 3 F0 output by Pitch command in SPTK. For unvoiced frames, SPTK outputs zero. The test data was prepared by overlapping noise at different SNRs. The noise was recorded in a car moving on an expressway with a fan at a medium level.

filter was designed in the spectrum domain for each frame as in Eq. (8), and the comb-filtering output was obtained using Eq. (9).

$$\begin{aligned} Wcomb_T(j) &= 1.0 && \text{if } T \text{ is unvoiced frame} \\ Wcomb_T(j) &= 1.0 && \text{else if } j \text{ is harmonic bin} \\ Wcomb_T(j) &= 0.01 && \text{else} \end{aligned} \quad (8)$$

$$z_T(j) = y_T(j) \cdot Wcomb_T(j) \quad (9)$$

For the combination of LPE and existing noise reduction algorithms, SS and ETSI Advanced Front-End (ES202-050) [3] were introduced in the evaluations. For SS processing, the first 0.1 second of each utterance was assumed to be a non-speech segment where the noise spectrum $N(j)$ could be estimated. The SS output was obtained as Eq. (10).

$$\begin{aligned} z_T(j) &= y_T(j) - \alpha \cdot N(j) && \text{if } y_T(j) - \alpha \cdot N(j) \geq \beta \cdot N(j) \\ z_T(j) &= \beta \cdot N(j) && \text{else} \end{aligned} \quad (10)$$

In this experiment, the subtraction weight α was set to 1.0, and the flooring coefficient β was set to 0.1.

3.3 Results of Standalone Test

Table 1 shows the resulting word accuracies for various environmental conditions. The baseline is the evaluation without using any speech enhancement or noise reduction algorithms. Table 1 also shows the estimated SNRs of the test data using the VAD (Voice Activity Detection) information came from the ETSI ES202-050. Note that the accuracy of SNR depends on the VAD information. Table 2 shows the estimated SNRs of the training data. We see CENSREC-3 trains an acoustic model at relatively better SNRs than for the test data. Therefore, speech enhancement and noise reduction are expected to help the test performance.

LPE enhances the local peaks considered to be harmonic structures. Therefore, a drawback is expected with LPE when the background noise contains music or speech from audio devices such as a radio, TV, or CD player, because the filter is designed to enhance that audio, too. This is a known restriction of LPE. Comb filtering shares this problem, and a multi-pitch tracker was proposed to address it [14]. In this paper, we accept this restriction and we focus only on the results of the "Audio off" cases. The restriction should not matter with current car navigation systems, because most of them are designed to disable audio on pushing a talk button. Also, we can expect an echo canceller to eliminate audio components before processing by LPE.

For the average "Audio off" case, LPE outperformed the baseline by 17.0% in error reduction. Most of the improvement was gained in very noisy conditions of "Fan high" and "Window open" conditions with error reductions of 14.8% and 23.7%, respectively. An advantage of LPE is that voiced speech immersed in heavy noise should be more distinct and distinguishable for decoding. Comb-filtering also improved the accuracy in these conditions. However, the improvement was smaller than LPE.

In relatively clean conditions such as "Normal" or "Fan low" at "Idling" or "Low speed", the accuracy of LPE was almost the same or slightly degraded from the baseline. However, the degree of loss was small enough for practical use. In contrast, comb-filtering shows noticeable degradation in these conditions, possibly caused by inaccurate F0 estimation and errors in the voiced/unvoiced detection.

Table 2 Estimated SNRs of CENSREC-3 training data. SNR was calculated for the baseline data after a 250 Hz high-pass filter.

Training Data	SNR (dB)
Idling	21.1
Driving	18.7

3.4 Results of Combination Test

LPE can be used in combination with existing noise reduction algorithms. In Table 3, SS and ETSI ES202-050 were introduced in the evaluations. Figure 4 shows the average word accuracies in combined "Audio off" cases. "SS+LPE" means LPE processed the output of SS. Since ETSI ES202-050 splits the 16-KHz input into a less-than-8-KHz part and an upper-8-KHz part, "ETSI+LPE" applied LPE only to the less-than-8-KHz part of the ETSI ES202-050 output.

The "SS+LPE" combination outperformed SS or LPE alone, as well as the baseline. It reduced the average error rate for the "Audio off" case by 27.3% from the baseline. Likewise, the "ETSI+LPE" combination showed the best performance, reducing the error rate by 69.2%.

Table 3 Word accuracy with existing noise reduction methods and the combinations of LPE.

CENSREC-3 (Condition 3)			Word Accuracy (%)			
			SS	SS + LPE	ETSI	ETSI + LPE
Idling	Audio off	Normal	99.8	99.0	100.0	100.0
		Hazard on	96.8	96.7	98.1	98.6
		Fan low	95.2	95.3	99.2	99.7
		Fan high	58.1	67.6	85.3	88.9
		Window open	90.4	93.8	97.2	98.0
	Audio on		74.8	61.4	89.5	82.6
Low speed	Audio off	Normal	98.4	97.5	99.7	99.7
		Fan low	94.6	94.2	97.8	98.7
		Fan high	66.9	74.3	87.9	91.5
		Window open	72.4	78.5	87.0	88.7
	Audio on		79.5	62.8	90.8	87.6
High speed	Audio off	Normal	97.8	95.9	98.1	98.8
		Fan low	91.7	91.6	96.7	97.6
		Fan high	61.3	69.6	88.4	88.1
		Window open	40.1	45.4	65.0	66.7
	Audio on		84.3	69.1	92.8	89.7
Average (ALL)			81.3	80.7	92.1	92.1
Average (Audio off)			**81.8**	**84.6**	**92.3**	**93.5**
Average (Audio on)			79.5	64.4	91.0	86.6
Average (Fan high)			62.1	70.5	87.2	89.5
Average (Window open)			67.6	72.6	83.1	84.5

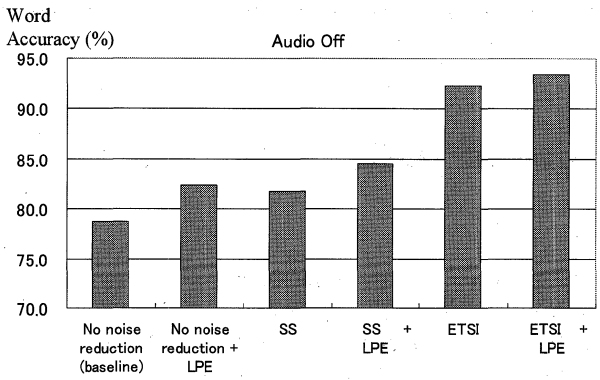


Fig. 4 Averaged word accuracy of "Audio off" cases for the combinations of noise reduction method and LPE.

4. Conclusion

We are proposing a new approach to speech enhancement to improve automatic speech recognition in very noisy conditions. It generates a filter to enhance the harmonic structure observed in the input spectrum, without relying on F0 estimation and voiced/unvoiced detection. Experiments using automatic speech recognition showed this method significantly improved the accuracy in very noisy conditions such as "Fan high" or "Window open". However, it showed some drawbacks in "Audio on" cases. This method can be combined with existing noise reduction algorithms such as SS and ETSI ES202-050 for further improvements.

References

[1] H. Saruwatari, K. Sawai, A. Lee, K. Shikano, A. Kaminuma, and M. Sakata, "Speech enhancement and recognition in car environment using blind source separation and subband elimination processing," Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation, pp.367–372, 2003.

[2] O. Ichikawa and M. Nishimura, "Simultaneous adaptation of echo cancellation and spectral subtraction for in-car speech recognition," IEICE Trans. Fundamentals, vol.E88-A, no.7, pp.1732–1738, July 2005.

[3] ETSI ES 202 050 v1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.

[5] H. Tolba and D. O'Shaughnessy, "Robust automatic continuous-speech recognition based on a voiced-unvoiced decision," Proc. ICSLP, paper 0342, 1998.

[6] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," Proc. ICASSP, vol.1, pp.125–128, 2001.

[7] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance spectrum based V/UV classification and F0 estimation," Proc. EuroSpeech, pp.2313–2316, 2003.

[8] L. Szymanski and M. Bouchard, "Comb filter decomposition for robust ASR," Proc. InterSpeech, pp.2645–2648, 2005.

[9] M. Sugiyama and K. Shikano, "LPC peak weighted spectral matching measures," ASJ Trans. of the Com. on Speech Res., S80-13, pp.101–108, 1980.

[10] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," Acoustical Society of America Journal, vol.116, no.4, p.2480, 2004.

[11] C. Huang, Y. Huang, F. Soong, and J. Zhou, "Weighted likelihood ratio (WLR) hidden Markov model for noisy speech recognition," Proc. ICASSP, vol.1, 2006.

[12] M. Fujimoto, S. Nakamura, K. Takeda, S. Kuroiwa, T. Yamada, N. Kitaoka, K. Yamamoto, M. Mizumachi, T. Nishiura, A. Sasou, C. Miyajima, and T. Endo, "CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework," Proc. International Workshop on Real-world Multimedia Corpora in Mobile Environments, RWCinME2005, pp.53–60, 2005.

[13] http://www.sp.nitech.ac.jp/~tokuda/SPTK/

[14] M. Wu, D. Wang, and G.J. Brown, "A multi-pitch tracking algorithm for noisy speech," Proc. ICASSP, vol.1, pp.369–372, 2002.